Data Mining Lab, Exercise 2

Team# 1: Poliakov Valerii, Holovnia Dmytro, Selvaraj Sinju
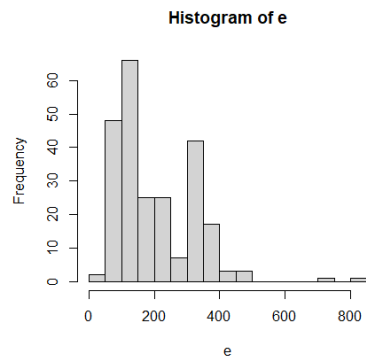
Dataset: cars 1.csv

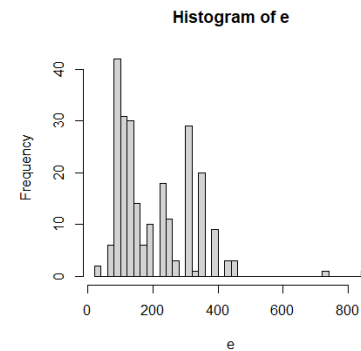## Task# 1

Identify the possible outliers using histograms and two-dimensional scatter plots.

d <- read.csv(file="cars 1.csv", header=TRUE, sep=",")
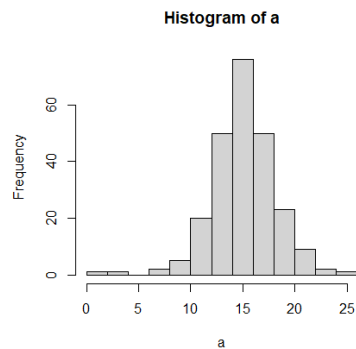*e <- d$engine.displacement*
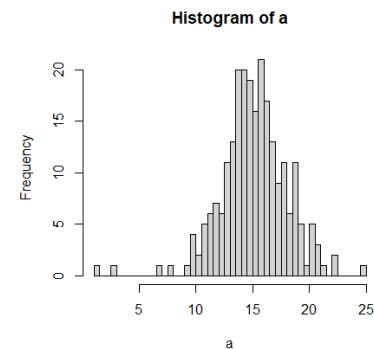*a <- d$acceleration*

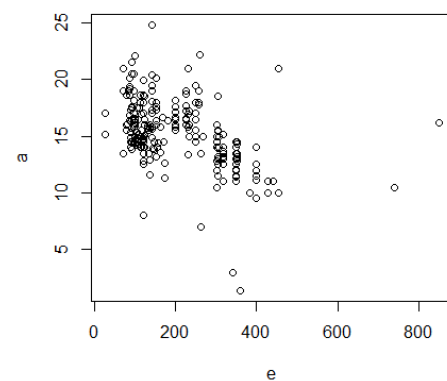*hist(e, breaks=15)*
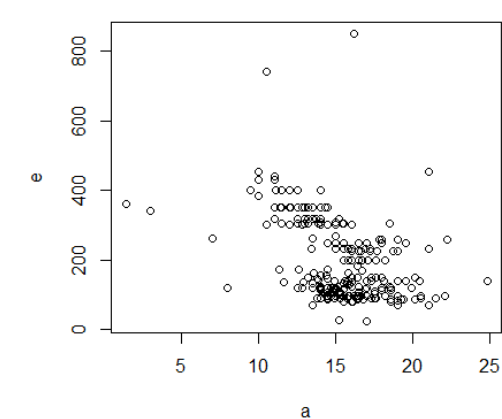

hist(e, breaks=50)

hist(a, breaks=15)

hist(a, breaks=50)

plot(e, a)

plot(a, e)

Values with e > 600(740, 850) and a < 5(3, 1.4) are the outliers. There are few more possible outliers that we need to check: a < 9(7, 8) and a > 24(24.8).

Task# 2

Verify that the values indicated in Task 1a and 1b are the outliers using:
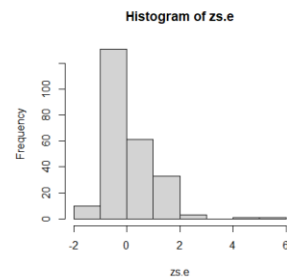
a) The Z-score method

zs.e <- (e-mean(e))/sd(e)

hist(zs.e)

zs.e.outliers <- e[(zs.e < (-3)) | (zs.e > 3)]

[1] 740 850

Values 740 and 850 are outliers by the z-score.

Histogram of zs.e

zs.a <- (a-mean(a))/sd(a)

hist(zs.a)

zs.a.outliers <- a[(zs.a < (-3)) | (zs.a > 3)]

[1] 3.0 1.4 24.8

Values 3, 1.4 and 24.8 are outliers by the z-score.

Histogram of zs.a
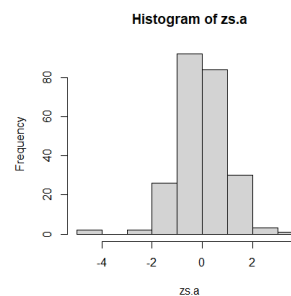
b) The IQR method

*iqr.e.outliers <- e[(e < quantile(e,0.25) - 1.5*IQR(e)) | (e > quantile(e,0.75) + 1.5*IQR(e))]*

*iqr.a.outliers <- a[(a < quantile(a,0.25) - 1.5*IQR(a)) | (a > quantile(a,0.75) + 1.5*IQR(a))]*

Values 740, 850, 3, 1.4, 22.1, 24.8, 22.2, 8 and 7 are outliers by the IQR method.

So, we can state that values 740, 850, 3 and 1.4 are verified outliers. Potential outlier 24.8 is confirmed. Potential outliers 7 and 8 are confirmed by the IQR method but not by Z-score method.

Task# 3

a) Which of the outliers detected by the numerical methods cannot be seen in the histogram and in the scatter plot?
- Values 22.1 and 22.2 outliers detected by the IQR method cannot be seen in the histogram and in the scatter plot.

b) Are there any outliers that can be clearly visible in the histogram or the scatter plot but are not indicated by the numerical methods?
- There is 1 value that is located at the maximum values of the range. There are no other values close to it, but it is not out of the range. So, this value cannot be detected by the range limits. Here we need to check number of values in the radius of the validated value. This method will show us that while value is in the range it is not intersecting with other values.

Task# 4

Investigate how the outliers affect the mean and median by doing the following:

a) Find the mean score and the median score with and without the outliers:

	e	a	e without outliers	a without outliers
Mean value	202.7063	15.23333	197.729	15.36936
Standard Deviation	118.9515	2.925629	106.1038	2.522276

b) State which measure, the mean or the median, the presence of the outliers affects more. Try to explain why.

Mean value is influenced by about 1.7%.
Standard Deviation is influenced by 12.7%.

The standard deviation is a measure of the amount of variation or dispersion of a set of values. So, when we remove the outliers found by methods like IQR or Z-score standard deviation value will go down.